Project Report

Enhanced Confidence for HER2 Status Predictions in Breast Cancer using Different Genomic Profiles

David Borer

(david.borer@pm.me)

June 2023

.....

Updated Report

.....

Supervisor
Eliezer M. Van Allen
eliezerm_vanallen@dfci.harvard.edu
Dana-Farber Cancer Institute
450 Brookline Avenue, D1230
Boston, MA 02115

Co-Supervisor
Olof Emanuelsson
olofem@kth.se
KTH Royal Institute of Technology
Tomtebodavägen 23A
171 65 Solna

Start and End Dates
Jan 17th, 2023 - Jun 5th, 2023

Table of Content

Abstract

Decisions in precision oncology are driven by extensive testing for many biomarkers. In breast cancer, HER2 is a well established biomarker that is traditionally tested for by immunohistochemistry (IHC). More recently, next-generation sequencing (NGS) has become widely used in cancer diagnostics. However, integrating the different data modalities from NGS such as the total copy number, allele-specific copy numbers and gene expression into current clinical practice is a difficult task. Here, we investigated the HER2 status concordance between different NGS data modalities with IHC in a breast cancer cohort (n = 687) and applied the learnings to a immuno-oncology (IO) cohort (n = 38). An accuracy of 96.07% was obtained by predicting every sample with an *ERBB2* total copy number higher or equal to 7 as positive. Integrating the gene expression as well lead to an accuracy of 96.36%. Logistic regression (LR) and random forest (RF) with the allele-specific copy numbers or all data modalities did result in very similar accuracies, though not improving the simple thresholding approach. The IO cohort was extremely limited in size, which prevented us from making any conclusions other than emphasizing the need for a bigger cohort. Our results confirmed the very high concordance between *ERBB2* amplification in breast cancer and the HER2 status by IHC. While we were not confident enough to recommend this integrative approach as a replacement for the IHC test, we recommend it as a complimentary tool.

Introduction

Dozens of molecular cancer properties are routinely used in the clinical setting to find the best treatments for cancer patients. Using a set of biomarkers to drive clinical treatment decisions is an important part of what is commonly referred to as precision oncology. Accordingly, testing for these molecular characteristics is a key step in the diagnostic process, especially with the higher availability of targeted therapies [1].

In breast cancer, the most often diagnosed cancer [2], molecular pathology has identified several biomarkers. Examples include mutations in *BRCA1* and *BRCA2* [3] and the expression levels of the estrogen receptor (ER), the progesteron receptor (PR) and the human epidermal growth factor receptor 2 (HER2, *ERBB2* gene). The latter is expressed in roughly 10 to 25% of all breast cancers worldwide [4, 5]. While HER2 amplification and overexpression is associated with poor prognosis [6], it also has predictive power for treatments targeting HER2 specifically [7]. Therapies include trastuzumab and pertuzumab, both therapeutic antibodies targeting HER2 directly, and over six other therapies that are fully or partially dependent on HER2 [8].

While HER2 is a robust and well-established biomarker, the biomarker landscape in immuno-oncology (IO) treatments is far-less so. IO treatments are based on the immune suppression mechanism of CTLA-4 and PD-L1/PD-1. Several therapeutic antibodies targeting CTLA-4 and PD-L1/PD-1 are available for the treatment of different cancers types, such as metastatic melanoma, non-small cell lung cancer (NSCLC) and renal cell carcinoma (RCC) [9]. Only roughly 20% of cancer patients enrolled in clinical trials can benefit from immunotherapies [10]. Predictive biomarker for anti-CTLA-4 treatments in metastatic melanoma are the mutational and neoantigen loads as well as expression of cytolytic markers. However, they can not fully explain response [11, 12]. On the other hand, PD-L1 (*CD274*) expression is associated with treatment efficacy in NSCLC [13]. Yet, PD-L1 expression does not guarantee response, and surprisingly, tumors without PD-L1 expression have been observed to respond to anti-PD-L1 therapy [9, 14]. The College of American Pathologists issued a seminal review in 2020 about the biomarker testing landscape for IO therapies [15]. Two biomarkers were approved by the U.S. Food and Drug Administration (FDA). First, the PD-L1 status by immunohistochemistry (IHC) and second, mismatch repair (MMR) and microsatellite instability (MSI), which are tested by IHC and PCR, respectively. Other biomarkers such as tumor mutational burden (TMB) and cancer neoantigens were deemed *emerging* and *early emerging*.

Traditionally, IHC is a commonly used technique to determine the status of a biomarker, such as HER2 and PD-L1. The American Society of Clinical Oncology set out clear guidelines to help clinicians in making the HER2 testing by IHC as reliable and reproducible as possible [16]. This generally is the case. Nevertheless, this does not mean that the test's accurcay is 100%. In fact, IHC is sensitive to tissue fixation, choice of antibodies and the determi-

nation of thresholds [17]. In the past, discordance between centralized and non-centralized laboratories could be as high as 20% [18, 19].

An alternative way to measure HER2 status is with next-generation sequencing (NGS). Depending on the specific sequencing technique, a NGS readout can provide information about gene sequences, gene copy numbers and gene expression [20]. The total gene copy number refers to the total number of copies of a gene present in a cell's nucleus. While a diploid cell has two copies per gene, tumor cells often undergo substantial genomic rearrangements, leading to many effects, such as copy number variations (CNVs). Total copy number is typically inferred from SNP array or whole exome sequencing (WES) data. Algorithms such as ABSOLUTE [21] and ASCAT [22] use additional information about heterozygous sites to compute the tumor purity and ploidy as well as the allele-specific copy numbers. The greater value of the two allele-specific copy numbers is commonly referred to as the major copy number, while the lesser value is referred to as the minor copy number. Gene expression can be obtained with NGS as well. RNA sequencing (RNA-seq) commonly refers to a technique that selectively captures mRNA molecules, which are then sequenced. For differential gene expression analysis, single gene expression levels are first quantified, followed by a normalization step for the dataset before batch-correction is applied [23].

Copy number profiles can be used as signatures for different cancer types [24, 25] and can have prognostic power in HER2-positive breast cancer [26]. Further, copy-number alterations can predict response to immune-checkpoint-blockade in gastrointestinal cancer [27]. Total gene copy numbers can also show concordance with the status determined by IHC, as it is the case for the HER2 status: detection of *ERBB2* amplification with a capture-based sequencing technique has shown 98.4% concordance to the status by IHC [28]. This has also been shown in esophageal cancer with a lower concordance (87.5%) [29]. On a pancancer level, higher copy number is strongly correlated with differential gene expression [30].

However, integrating total and allele-specific copy number profiles as well as gene expression profiles into current clinical practice is a difficult task. As described above, concordance between these data modalities can be very high. Based on this, we hypothesize that these data modalities, besides showing concordance, can be used to increase confidence for clinically relevant genomic events. Specifically, we asked whether using allele-specific copy numbers and gene expression could increase the confidence compared to using total copy numbers only. To test this, we predicted the HER2 status by IHC from total gene copy numbers, allele-specific copy numbers and gene expression in a set of 687 breast cancer samples. Additionally, we performed a similar analysis in 38 samples that were previously treated with immuno-oncology therapies, where we tried to predict the response from the total copy numbers of PD-L1/PD-1 and CTLA-4.

Methods

Data source

Three sources have been used for this work, where the first two sources where used for the breast cancer cohort.

(1) Total copy number, allele-specific copy number and gene expression where retrieved from the NIH Genomic Data Commons website[1] relating to the publication from Hoadley *et al.* [31]. From there, the ABSOLUTE-annotated seg file and the RNA batch corrected matrix have been downloaded. In this dataset, 2216 samples were from breast cancer, of which 1020 had total copy numbers, allele-specific copy numbers and gene expression. Sample data was matched using the sample ID.

(2) The HER2 status by IHC was retrieved from supplementary table 1 from the breast cancer publications from the TCGA (n = 825) [32]. The row `HER2 Final Status` contained the HER2 status by IHC for 776 samples. Equivocal samples were removed (n = 9), resulting in 766 samples used. HER2 status by IHC was matched with the genomic data using the patient ID.

694 samples had all data modalities after merging the two sources. 9 metastatic samples as well as one duplicate were removed, resulting in 687 samples that were used in the analysis. An overview of this cohort is provided in figure 1.

(3) The IO cohort was obtained internally, but the data is available in the supplementary information of Miao *et al.* [33]. There were a total of 62 samples, of which 44 could be classified as responders or non-responders based on the RECIST criteria [34]. The classification as responders and non-responders was done in the followng way: *complete response* (CR) and *partial response* (PR) were deemed responders, *progressive disease* (PD) as non-responders while *stable disease* (SD) samples were not used. From these 44, there were 28 samples that received anti-CTLA-4 therapies, 10 that were treated with anti-PD-L1/PD-1 therapies and 6 that received both therapies. For the analysis, the 6 double-treated samples were not used.

Software, code and hardware

All analysis have been done in JupyterLab 3.5.3 with Python 3.9.7. The main used packages were matplotlib 3.6.2, numpy 1.21.5, pandas 1.4.4, plotnine 0.10.1, scikit-learn 1.2.1, scipy 1.10.0 and seaborn 0.12.2. The code is available on GitHub[2]. A personal laptop has been used as hardware, running Windows 10 with 16 GB RAM and 2.70 GHz dual core Intel i7 CPU. Some analysis such as 10'000 iterations of logistic regression were running for low single digit minutes while 10'000 iterations of random forest were running for low double digit minutes depending on other use of the computer.

[1] URL: https://gdc.cancer.gov/about-data/publications/PanCan-CellOfOrigin
[2] URL: https://github.com/davidborer/master-thesis-project.git

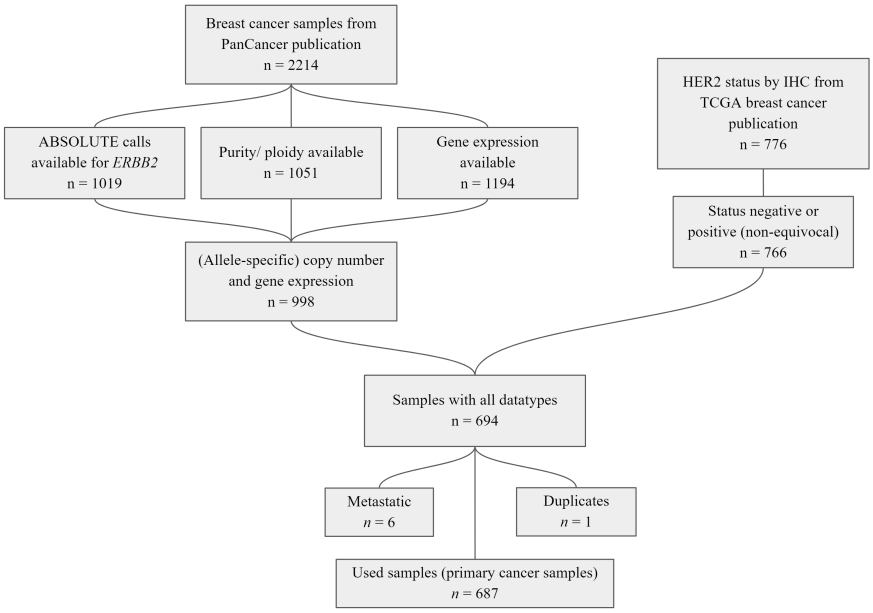


Figure 1: The PanCancer publication from Hoadley *et al.* [31] contained 2214 breast cancer samples and the TCGA publication for breast cancer contained 776 samples with the HER2 status by IHC [32]. These samples were matched using the sample ID for the total gene copy number, allele-specific copy number and gene expression, and then using the patient ID for matching with the status by IHC. By keeping only primary tumor samples and removing one duplicate, the final set contained 687 samples.

Threshold setting

The total gene copy numbers and the gene expression have been used as single modality or in combination to predict the HER2 status by IHC. The thresholding approach consistent in simply setting a threshold, meaning that all samples with value equal or greater to that specific threshold were predicted as positive. As an example, a threshold of 5 for the total gene copy number would mean that any sample that has a total copy number of 5 or more was predicted as being positive.

The thresholding approach has been applied to the whole cohort as well as by splitting the data in 80:20 where 80% were used to find the best threshold and 20% to assess the predictions made with this threshold. More about the splitting is detailed in the last paragraph of the next subsection (Logistic regression).

Logistic regression

Logistic regression is a simple supervised learning algorithm that is used to predict binary outcomes by assigning a probability. If the two outcome labels are *negative* and *positive*, a sample with a probability over 50% could be predicted as *positive* (or vice-versa).

The formula below is the logistic function, where β_0 is the intercept and β_1 the coefficient of the first feature. If there are more than one feature, every additional feature is reflected by an additional β-term.

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_m)}}$$

While the main function of logistic regression is to make predictions, the coefficients can be used to asses the importance of the feature for the outcome. However, the βs can only be used if the input feature follow a Gaussian distribution and are normalized (e.g. mean = 0 and standard deviation = 1).

The coefficients and the intercept can be extracted from the model in scikit and used to calculate the probability of any input values. In the case of having two input features, we can express the value of a feature based on the other feature(s) for the case in which logistic regression would return a 50% probability. Applying this to the case where the features are the allele-specific copy numbers with the major copy number on the x-axis, a function $f(x) = y$ would, given the major copy number, predict the minor copy number at decision boundary of the logistic regression model. From the logistic function, we can formulate the function below, where x_1 is the major copy number and y the minor copy number.

$$y = -\frac{\beta_0 + \beta_1 x_1}{\beta_2}$$

The scikit class `StratifiedShuffleSplit` has been used to split the data in training and test sets. This class ensure that the label distribution is maintained in the split (in our case the ratio of positive versus negative samples). A split of 80:20 has been chosen. No random seed has been set, meaning that the split was different for every iteration. The scikit-learn class for logistic regression (`sklearn.linear_model.LogisticRegression`) has been used for training a model and make the predictions, which where assessed with scikit-learn accuracy and F1 scores. The performance was estimated by iterating over the training, prediction and assessment steps many times.

Random forest

Random forest is a simple machine learning algorithm that is based on many decision trees that are build with random subset of the dataset. These decision trees are used for making predictions, with the final result being the most often predicted outcome. The randomness in the tree building usually prevents overfitting, but it can still happen. Overall, random forest is a more complex

algorithm than logistic regression and is therefore able to better represent more complex data. However, its interpretability is generally less good.

Random forest has been implemented in the same way as logistic regression was, meaning that the splitting was the same as well as the iterating process to estimate its performance. The scikit class for random forest is `sklearn.ensemble.RandomForestClassifier`. The importances of the input features can be accessed using `.feature_importances_`, which results in a score giving the relative importance from 0 to 1.

Performance assessment

The performance of the different methods was assessed using the accuracy and the F1 score. The accuracy expresses the ratio of correctly predicted samples by dividing the number of correctly predicted samples by the total number of samples. Multiplying this number by 100 results in the accuracy in percents, which is a simple and easy to understand score. It can also be calculated using the TP (true positive), FP (false positive), TN (true negative) and FN (false negative) rates.

$$\text{accuracy} = \frac{\text{number of correctly predicted samples}}{\text{total number of samples}} = \frac{TP + TN}{TP + FP + TN + FN}$$

However, the accuracy is not correcting for a possible sample label imbalance. In a sample population with 80% negative samples and only 20% positive samples, predicting all samples as negative will result in a 80% accuracy. This does not accurately reflect the performance of the made predictions. The F1 score takes this inbalance into account by using the precision and the recall. The precision is a measure of how precise the positive predictions are, meaning that predicting too many samples as positive decreases the precision. On the other side, the recall is a measure that express how many true positive were actually found, meaning that it is a measure of sensitivity for the positive samples. The F1 score combines them and thus provides a sensible scoring system that is also sensitive for inbalanced sample populations. The best F1 score is 1.

$$\text{F1 score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

where

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

Results

687 samples containing total gene copy numbers, allele-specific copy numbers and gene expression for the gene *ERBB2* have been retrieved based on the publication from Hoadley *et al.* [31]. The HER2 status by IHC, our ground-truth, was retrieved from a supplementary table from a TCGA publication [32]. An overview of these samples is shown in figure 2. The dataset contained 84 HER2-positive and 603 HER2-negative samples according to the status by IHC (12.2% HER2-positive samples).

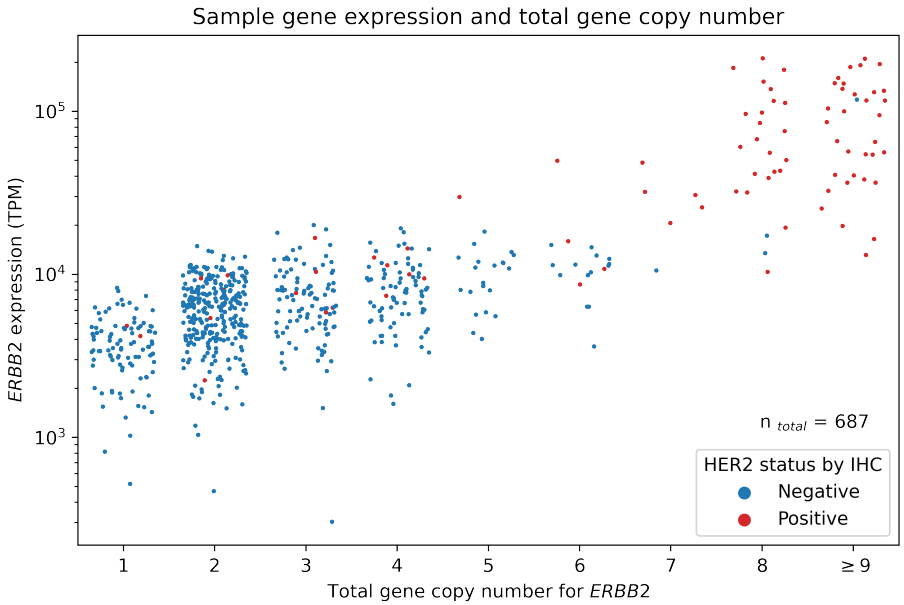


Figure 2: 687 samples with gene expression and gene total copy number, colored by their respective HER2 status by IHC. This dataset was used through all analysis for breast cancer. Copy numbers (CN) of 9 and more include 4 samples with 10 copies, 2 samples with 11 copies, 1 sample with 12 copies and 1 sample with 14 copies. The copy numbers were all integers; for better visualization, a certain spread has been added on the x axis.

Total copy numbers

The total copy number for *ERBB2* was used to predict the HER2 status, with the status determined by IHC as being our ground-truth. By defining a threshold of 7 copies, meaning that a sample having 7 or more gene copies for *ERBB2* was classified as HER2-positive, an accuracy of 96.07% was achieved. As the data is heavily skewed towards HER2-negative samples, thresholds greater than 7 do not lead to an accuracy below 87.8% since 12.2% are HER2-positive (see suppl. figure 1). However, the F1-score, which accounts equally for false positives and

for false negatives, clearly displays a threshold of 7 being optimal (see suppl. figure 2). The confusion matrix (suppl. figure 3) shows that 599 out of 603 HER2-negative samples were correctly predicted (99.3%) whereas only 61 of 84 HER2-positive samples were correctly predicted (72.6%). This indicated a sensitivity problem.

Applying the thresholding approach to the whole dataset beared the potential of overfitting. Therefore, the thresholding approach has been used with a stratified 80:20 split over 10'000 iterations, meaning that 80% of the data have been used to find the best threshold while 20% have been used to calculate the accuracy of the predictions made with the calculated threshold. The randomness in the train sets lead to 307 iterations (3.07%) finding 6 as being the optimal threshold (all other iterations determined 7 as optimal) and the randomness in the test sets added variation as well in the calculated accuracies. The average accuracy after 10'000 iterations was 95.94%.

After 10'000 iterations of data splitting, logistic regression model training and predictions with the total copy number, an average accuracy of 94.75% was obtained. Using a random forest model, the average accuracy was 96.03%.

Gene expression

The *ERBB2* gene expression was used as well to predict the HER2 status. A threshold was defined similarly to the total gene copy number, resulting in a threshold of 19'200 TPM as having the highest accuracy and F1 score (see suppl. figures 4 and 5). With an accuracy of 96.22%, it was slightly higher than the predictions made with the total copy numbers. Indeed, only 2 of 603 HER2-negative samples were wrongly predicted as positive (99.7% correct), but only 60 of 84 HER2-positive samples were correctly predicted (71.4%) (see confusion matrix in suppl. figure 6). This also suggested a sensitivity issue (more false negatives than false positives). Figure 2 confirmed that there is a number of HER2-positive samples with low copy number and low gene expression.

Thresholding using the gene expression was also carried out with a 80:20 split, resulting in an average accuracy of 95.63% after 10'000 iterations, whereas random forest resulted in 93.30% accuracy. Logistic regression did not perform well (49.74%), which has not been investigated further.

Total copy number and gene expression

The integration of total *ERBB2* copy number or *ERBB2* gene expression as a single data modality led to 96.07% and 96.22% overall accuracies. Using the thresholding approach, both data modalities were integrated in two different ways. The first was a `AND` combination, where a certain threshold needed to be met for the total gene copy number *and* the gene expression for a sample to be predicted as positive. The second approach was a `OR` combination, where a sample was predicted positive when either the total copy number *or* the gene expression were above a certain threshold. To find the best performing thresh-

olds, two heatmaps of the F1 scores with different threshold combinations were made for both the `AND` and `OR` combinations (see suppl. figures 7 and 8). They indicated that for a `AND` combination, a threshold of 6 total copies and 15'500 TPM were best, resulting in an accuracy of 96.36%. Figure 3 shows the thresholds and the resulting sample population that is predicted as positive. For the `OR` combination, a threshold of 7 copies and 21'000 TPM lead to an accuracy of 96.36% as well. The selected samples are shown in figure 4.

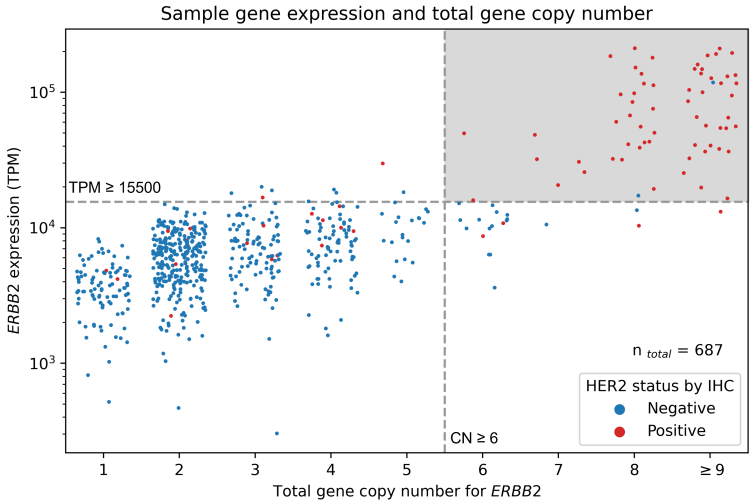


Figure 3: Combining both data modalities by predicting a sample as positive only if both modalities are above certain thresholds (`AND`) lead to lower threshold-setting. The accuracy of the shown thresholding is 96.36%.

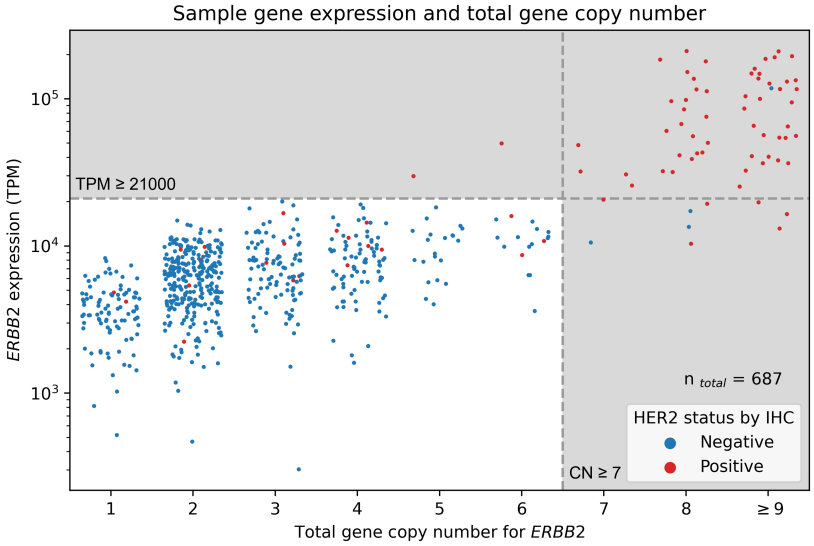


Figure 4: Combining both data modalities by predicting a sample as positive if any of both modalities is above a certain threshold. The `OR` combination lead to higher threshold-setting. The accuracy of the shown thresholding is 96.36%.

Allele-specific copy numbers

Next, the allele-specific copy numbers of *ERBB2* were used to predict the HER2 status. Figure 5 provides an overview of the distribution of the allele-specific copy numbers colored by their HER2 status (IHC ground-truth). Many samples had a different HER2 status while displaying the same combination of major and minor copy numbers. Based on figure 5, there are different ways a thresholding approach could be used.

A first option was to classify every samples as positive if the sum of the major and minor copy numbers are greater or equal than 7. However, this is the same as using the total copy number.

A second option was to classify every samples based on the major or minor copy number only. Figure 5 suggested to use the major copy number for this task. Using the same procedure as with the total copy number previously, a threshold of 6 copies was found to be best. It resulted in an accuracy of 95.63%.

A third option was to make a complicated thresholding system to optimally separate the samples. This has been deemed unpractical. However, a hypothetical separation line can be drawn, assigning every copy number combination to its most represented status. Assuming that the combination [minor=2, major=6] would have been classified as positive by this hypothetical line, a theoretical accuracy of 96.36% resulted. (The combination [minor=0, major=7] did not influence the outcome since there were one positive and one negative sample.)

Using logistic regression with a stratified 80:20 split and the major and minor copy number as features, an average accuracy of 95.77% has been achieved after 10'000 iterations. Figure 6, boxplots A and B show the distribution of the accuracy and F1 scores. The accuracy score spread from roughly 90% to 100% was due to the randomness in the split of the data for the training and test set. Indeed, the distributions of the copy number from the 500 iterations that led to the highest F1 score and the 500 iterations that led to the lowest F1 score are different (see suppl. figures 10 and 11). This means that higher F1 score were achieved by having a test set that contained more unambiguous samples compared to the test sets that led to the lower F1 scores.

Similarly to logistic regression, random forest has been used to classify the samples based on their major and minor copy numbers. Figure 6, boxplots C and D show the distribution of the accuracy and F1 scores for 10'000 iterations. Random forest resulted in an average accuracy of 95.83%.

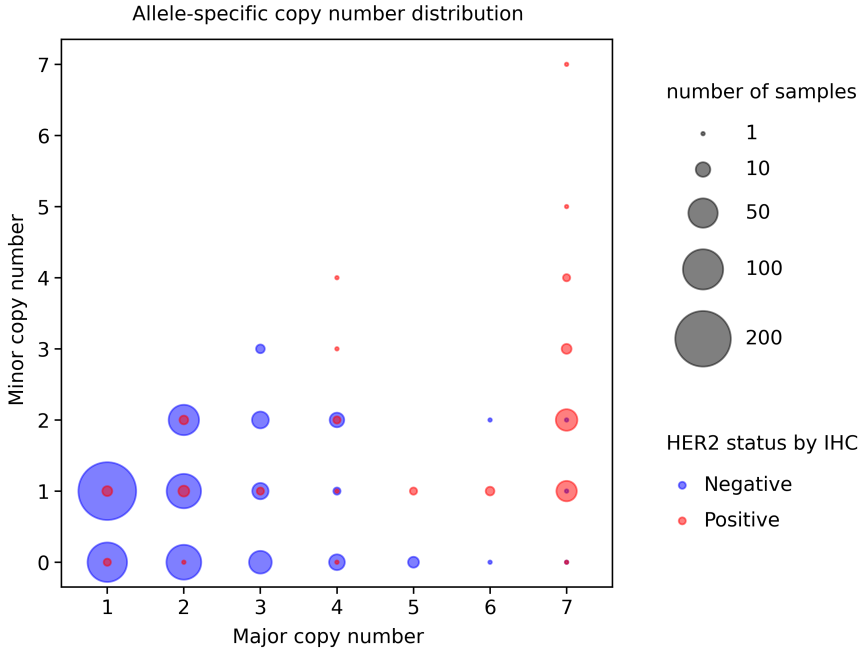


Figure 5: Allele-specific copy number of *ERBB2* for all samples.

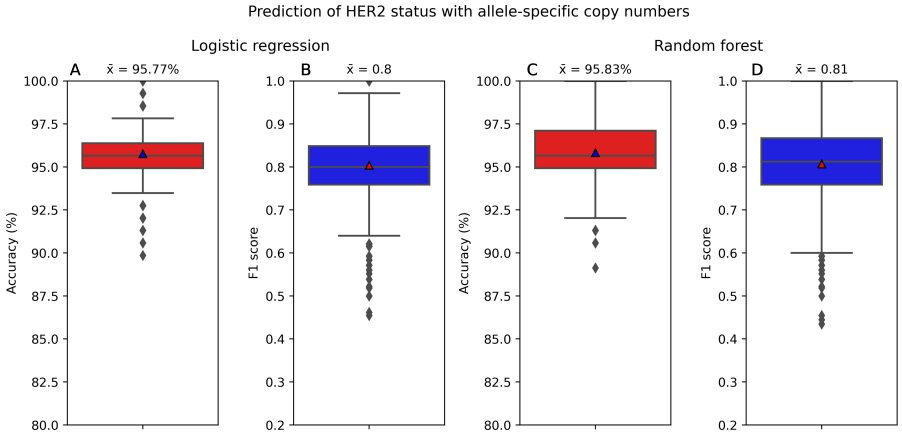


Figure 6: Training a logistic regression model and a random forest model with the allele-specific copy numbers for *ERBB2* using 687 samples in total. For each model type, 10'000 separate iterations of data splitting (80:20), training, predictions and assessment (accuracy and F1 score) have been made. Triangles in boxplots are the means.

All data modalities and feature importance

The total copy numbers, allele-specific copy numbers and gene expression for *ERBB2* were used as input features for logistic regression and random forest. This resulted in a 95.96% accuracy with logistic regression and 94.04% with random forest (see suppl. figure 9). To interpret the coefficients from logistic regression, all 4 features have been normalized before the training by subtracting their respective mean and dividing them by their respective standard deviation. However, the distribution of the features were not Gaussian, which prevented us from interpreting them. With regards to random forest, the features importance were 0.54 for the gene expression, 0.26 for the total copy number, 0.17 for the major copy number and 0.03 for the minor copy number. When not using the gene expression, the total copy number and the major copy number were similarly important (both around 0.45), while the minor copy number was less important (0.10). If using only the allele-specific copy number, the major copy number (0.85) is more important than the minor copy number (0.15).

IO cohort

The immuno-oncology (IO) cohort contained 28 samples that either responded or not to an anti-CTLA-4 treatment and 10 samples that either responded or not an anti-PD-L1/PD-1 treatment. The samples had all three data modalities: the total gene copy numbers, the allele-specific copy numbers and the gene expression. The response was assessed using the RECIST criteria [34]. The

anti-CTLA-4 IO cohort had 8 responders (29%) and the anti-PD-L1/PD-1 IO cohort had 5 responders (50%).

Using a thresholding approach, it was not possible to find a sensible threshold of the total gene copy number for *CTLA4* to predict the response to the anti-CTLA-4 treatment, since the best accuracy (∼70%) was achieved using a threshold that predicts every samples as non-responder (≥ 6), whereas the best F1-score (∼0.45) has been achieved with a threshold that predicts every sample as responder. Using logistic regression with the allele-specific copy number as features, an accuracy of roughly 70% has been reached while the F1 score was equal to 0. Gene expression did not improve these results. Figure 7 shows the sample distribution according to the total gene copy number and gene expression.

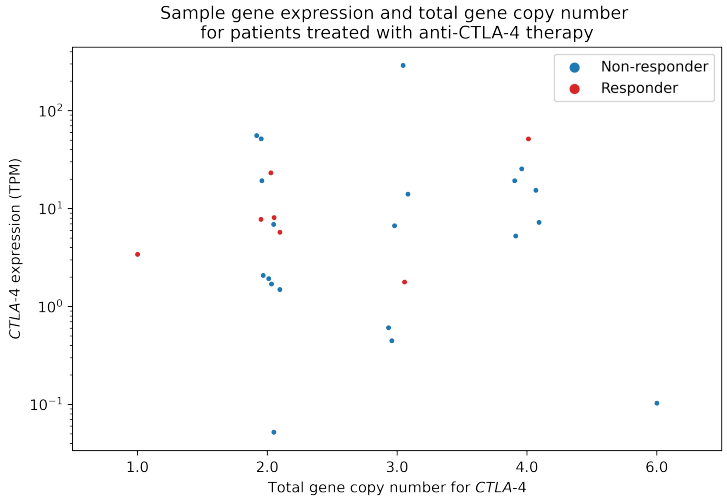


Figure 7: Overview of samples from the IO cohort subset that have been treated with anti-CTLA-4 therapy. The shown total copy numbers and gene expression are for CTLA-4. The total number of samples is 28.

Similarly to the anti-CTLA-4 subcohort, it was not possible to find a sensible way to integrate the different data types in the subcohort treated with anti-PD-L1/PD-1 therapies. The thresholding approach with the total copy number lead to an accuracy of 70% with a threshold of 2 total copies for *CD274* (PD-L1). The F1 score was between 0.7 and 0.8. With this threshold, 5 samples were correctly predicted as responders, 2 correctly as non-responders while 3 samples were wrongly predicted as responders. Logistic regression using the allele-specific copy number for *CD274* led to an average accuracy of ∼70%, with an average F1 score of ∼0.65. However, both score were distributed from 0 to 100% (or 1), which is due to the fact that the test set with a 80:20 split contained only 2 samples. All samples are shown in figure 8. Gene expression for *CD274* did not improve these results. The samples that received anti-PD-

L1/PD-1 treatment were also analysed regarding the different data modalities for *PDCD1* (PD-1). Setting a threshold of 2 copies for the total copy number resulted in an accuracy of 60% (F1 score = ∼0.7). Indeed, 5 samples were correctly predicted as responders, 1 sample was correctly predicted as non-responders and 4 samples were wrongly predicted as responders. Logistic regression with the allele-specific copy numbers for *PDCD1* resulted in an average accuracy of ∼50% (F1 score 0.5). Here again, a 80:20 split leads to a test set with 2 samples. The samples are shown in figure 9.

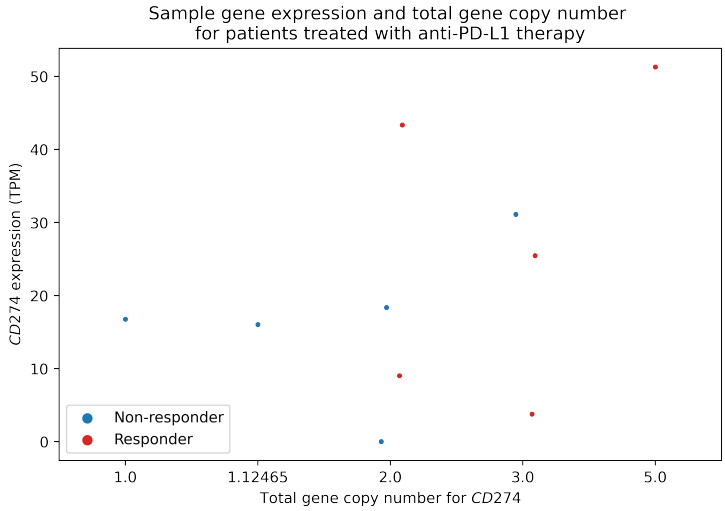


Figure 8: Overview of samples from the IO cohort subset that have been treated with anti-PD-L1/PD-1 therapies. The shown total copy numbers and gene expression are for PD-L1 (*CD274*). The total number of samples is 10. The y-axis is not logarithmic since one sample had a TPM value of 0.

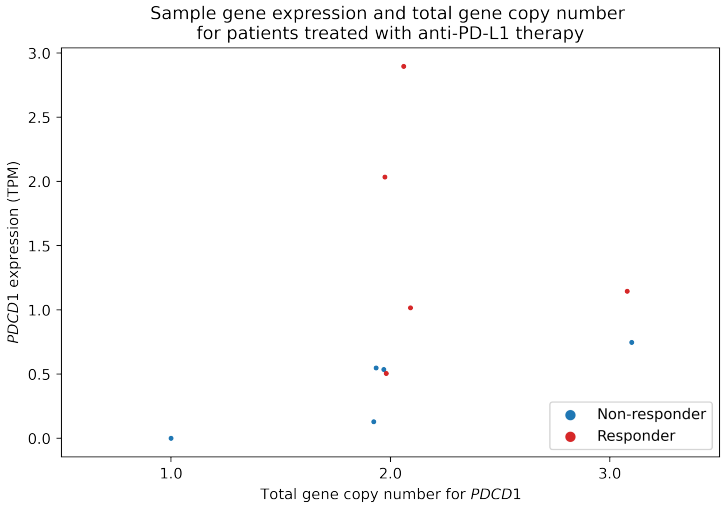


Figure 9: Overview of samples from the IO cohort subset that have been treated with anti-PD-L1/PD-1 therapies. The shown total copy numbers and gene expression are for PD-1 (*PDCD1*).

Discussion

The HER2 status in breast cancer was predicted using the total *ERBB2* copy numbers, the allele-specific *ERBB2* copy numbers and the *ERBB2* expression data. The cohort contained 687 samples of which 12.2% were HER2-positive according to IHC.

A threshold of ≥ 7 copies for the total gene copy number, meaning that every sample with 7 or more copies for the gene *ERBB2* was predicted as being HER2-positive, lead to an accuracy of 96.07% (measured across the whole cohort). Using the gene expression as single modality, a threshold of 19'200 TPM lead to an accuracy of 96.22%. Integrating both modalities with a thresholding approach resulted in an accuracy of 96.36%. This simple thresholding approach indicated a clear concordance between the total copy number and the HER2 status by IHC. This was expected since HER2 in breast cancer is a well characterized biomarker, as Ross *et al.* demonstrated using a targeted cancer gene sequencing assay. They determined the copy number amplification in 213 samples and found a 98.1% concordance between *ERBB2* amplification and the status by IHC/FISH (fluorescent *in situ* hybridization) [28].

Setting a threshold for the whole cohort and measuring the resulting prediction accuracy with the whole cohort exposed us to a risk of overfitting. Subsequently, a balanced splitting was made, where 80% of the data was used to determine the best threshold and 20% to measure the accuracy. This confirmed that a total copy number of 7 is best. Nevertheless, the randomness in the splitting meant that the train and test set were different for every iteration, leading

to an average accuracy of 95.94% after 10'000 iterations. Although it is very similar to the result with the total cohort, it highlights the importance of not using the whole dataset for training and testing.

Logistic regression and random forest are simple prediction algorithms for categorical outcomes. Logistic regression calculates coefficients for every input feature to make a binary decision based on an assigned probability. On the other hand, random forest uses many decision trees to independently classify the input and then averages the result to a final decision. Using logistic regression with the total copy number as input (1 feature), an average accuracy of 94.75% was achieved after 10'000 iterations, while random forest resulted in an accuracy of 96.03%. This highlights that random forest is well suited for this task, since a decision tree is comparable to a thresholding approach. With logistic regression, an accuracy of 95.77% was achieved using the allele-specific copy numbers (2 features) and 95.96% was achieved using all data modalities (4 features). These accuracies are the average of 10'000 iterations where a different balanced 80:20 split was made for every iteration. While the average performance is extremely similar to the thresholding approach, many iterations had a higher accuracy than the thresholding approach. This was explained by the differences in the test sets, where the top 5% iterations had less ambiguous test samples than the bottom 5%. An ambiguous sample here was defined as one that exhibits a HER2 status by IHC opposite to the majority of samples with the same combination of allele-specific copy numbers. Using random forest to predict the HER2 status with the allele-specific copy number only resulted in an accuracy of 95.83%. This result is slightly better than logistic regression, but due to the randomness in the 10'000 iterations, it would be wrong to say that random forest outperformed logistic regression. Using all data modalities, random forest resulted in an average accuracy of 94.04%.

Logistic regression is generally seen as more interpretable than random forest since the coefficients can be used to determine the relative importance of the input features. Therefore, we normalized the input features in the hope of using the coefficients to draw conclusion on each ones importance for determining the outcome. However, the features were not normally distributed, which was expected to some extend, especially for the copy numbers. Most samples had a total copy number of 2, and while a copy number smaller than 0 is impossible, one sample was as high as 14 in our case. In random forest, the relative importances of the features can be extracted. This indicated that, when using all 4 modalities, the gene expression was the most important. Interestingly, using the total copy number as only input, random forest provided the best average accuracy, which would suggest that adding other modalities such as the allele-specific copy number and the gene expression had a negative effect on the accuracy. However, since all accuracies were very similar, conclusions are very preliminary at this point.

While the concordance between the HER2 status by IHC and the different genomic data types was generally very high, only 61 of 84 HER2-positive samples (72.6%) were correctly predicted using the total copy number (while 99.3%

of the HER2-negative samples were correctly predicted). Similarly, Ross *et al.* were able to predict all HER2-negative samples (n = 155) correctly, but missed 4 of 58 HER2-positive samples (93.1% positive concordance). Our cohort indeed contained 5 HER2-positive samples with only 2 *ERBB2* copies and 2 with only one single copy (see figure 2). This suggested that false-negatives are generally a more important problem compared to false-positives, indicating a discordance between the protein levels and the gene expression or copy number. Memon *et al.* reported up to 12.5% discordance between IHC-positive and ISH-negative samples between 2001 and 2007 (ISH: *in situ* hybridization) [35]. This discrepancy has decreased over the years and was at 5.8% after 2019. As ISH methods measure gene amplification, this confirmed that there are cases of IHC positive test results without detecting gene amplification.

Interestingly, previous publication also highlighted the opposite. Luoh *et al.* reported in 2013 several samples with gene copy increases without a higher level of expression [36]. More recently, Boichard *et al.* reported in 2021 a subset of samples with highly amplified *ERBB2*, defined as 6 or more copies, not having high mRNA levels for *ERBB2*. They found that focal amplifications[3] often are silenced compared to non-focal amplifications [37].

Some limitation of this work might also be the cause for the discordance of roughly 4% of the samples, next to possible biological reasons. First, tumor heterogeneity could lead to discordance between the HER2 result by IHC and the NGS result, since it has not been performed on the exact same tissue. Second, the data for the HER2 status has been matched to the genomic data by using the patient ID. We assumed that the samples used for the status by IHC determination were primary samples and accordingly only used genomic samples from primary sites. This is a possible cause for discordance, since it has been reported that metastatic lesions can in rare cases have a different HER2 status (either become HER2-positive or negative depending on the primary site's status) [38]. Nevertheless, from 694 samples, only 6 were from a metastatic site (less than 1%). A third limitation of this work is the IHC data, since we did not have the stages (0, 1+, 2+ or 3+). The status was either positive, negative, or equivocal (9 equivocal out of 776 samples). In general, the status by IHC is known to not be 100% reliable. The IHC assay is sensible to tissue fixation, choice of antibodies and the determination thresholds [17] and the disparity between centralized versus non-centralized laboratories could be as high as 20% [18, 19]. Nevertheless, the guidelines for HER2 testing have been constantly updated and HER2 is generally a well-established assay [39]. Last but not least, a fourth limitations is that the genomic data types were not compared to the treatment response, which is what is ultimately important for the treatment decisions.

In a second part of this work, we analysed 28 samples from patients treated with an anti-CTLA-4 therapy and 10 samples from patients treated

[3] A focal amplification refers to a local amplification affecting one or a few genes. There is no standard definition, but it is usually referred to as being smaller than 10 Megabases. On the other hand, a non-focal amplification can affect whole chromosomes.

with an anti-PD-L1/PD-1 therapy. The cancer types varied, but were mainly melanoma. The response to anti-CTLA-4 treatment could not be predicted with any of the data modalities. The sample population in figure 7 suggested that having more than two *CTLA4* copies was associated with non-response. However, a copy number of 2 was not associated with any outcome. This is in contrast to a study from 2017 suggesting that a copy loss was associated with non-response [40], which the authors argued was because immune pathways would be less active. The response to anti-PD-L1/PD-1 treatment could have been predicted with 70% accuracy using a threshold of 2 or more copies for *CD274* (PD-L1). However, it was clear that any conclusion regarding the IO cohort are not possible since it is underpowered. This was firstly due to our small samples size, with only 28 samples for the anti-CTLA-4 cohort and 10 samples for the anti-PD-L1/PD-1 cohort. Results from logistic regression or random forest therefore were mostly meaningless. Despite this, our data was in line with Hong *et al.* stating that a PD-L1 copy loss is associated with non-response [41]. In general, the ground-truth of this dataset (response) is much more noisy than the endpoints of the breast cancer cohort (HER2 status by IHC). Response to treatment is not only influenced by the treatment itself, but by many other co-factors as well.

Taking a step back, the HER2-status by IHC is highly concordant with genetic analysis. This was expected since the HER2 status in breast cancer is a well-established biomarker. A HER2-positive status is a prognostic factor which is associated with a more aggressive disease. At the same time, it is a strongly predictive factor for the likelihood of response to a HER2 therapy, which means that sensitivity is very important. With the NGS approach missing out HER2-positive samples that have low copy number and low gene expression, it is hard to say if NGS alone is enough for predicting the HER2 status. However, our cohort had 4 HER2-negative samples with a total copy number higher than 7, of which one had a copy number of 9. It has been reported that a IHC-negative but NGS-positive patient for HER2 eventually respond completely to anti-HER2 therapy [42]. Therefore, NGS could be used to screen for potentially missed HER2-positive samples. Additionally, it was interesting to see that the total copy number provided as much predictive power as all other used modalities combined. Lastly, random forest seemed to work best for integrating the total copy number.

Future Work

This work suggests that the total copy number can help refine traditional diagnostics in the case of gain-of-function events. Important learnings can be drawn from the concordance between the NGS data and IHC testing generally. Indeed, in the case of HER2, we found that the total copy number alone is very informative. In breast cancer, NGS could be used to ensure no HER2-positive patient stays undetected. For PD-L1, no conclusion was made due to the small cohort size. Total copy number for PD-L1 are still interesting since Hong *et al.* reported that a copy loss of PD-L1 identified the worst responsive group, while IHC for PD-L1 was able to detect the most responsive group of patients for anti-PD-L1 therapy [41]. This loss-of-function event is in contrast to the gain-of-function event of HER2.

In biomarker discovery, IHC testing is not always available. We think that any case with such a high concordance between total copy number, allele-specific copy number and gene expression as in breast cancer is suggesting it can be used as a biomarker.

In the future, including other cohorts, different biomarkers and alternative datatypes could help refine biomarker accuracy. First, having the IHC grades would probably help understand the relationship between the copy numbers and the status by IHC and would allow for more refined predictions using random forest and logistic regression. Second, since HER2 is expressed in many other cancers as well [43], analysing the concordance there would be important as well. Third, integrating more data types, such as structural variants, amplification focality and copy numbers of other genes could lead to a higher accuracy. This would require additional raw data though, such as whole exome or genome sequencing. Last but not least, a greater cohort size is required for the IO analysis. However, finding biomarker for response to IO treatment is a very difficult task. Other biomarker should be investigated before tackling such hard cases. An example could be the androgen receptor in prostate cancer.

Acknowledgements

First and foremost, I would like to thank Brendan Reardon, who was my mentor throughout this project. His scientific guidance and constant support have been invaluable. I would like to thank Eliezer Van Allen as well, who accepted me to join his research group. I am grateful for his scientific expertise and regular feedback. Additionally, I would like to thank Jihye Park and everyone else in the Van Allen lab for the insightful conversations and the fun. Further, I would like to thank my co-supervisor Olof Emanuelsson for his important feedback on my report. Finally, I would like to thank the Fulbright Program for the financial and administrative support, without whom this would not have been possible.

Ethical Reflection

There are several points to consider while reflecting on the ethics of this project.

(1) Data privacy and confidentiality. This project used genomic data and clinical metadata that are sensitive information related to many patients. This information should be anonymized. In this work, the data was obtained from a public repository and internally. Both dataset handled in this work did not contain enough information to identify a patient, since we did not use any DNA sequences, but rather copy numbers and RNA counts only.

(2) Informed consent. This data has been obtained from patient with informed consent. The TCGA states clearly that informed consent has been obtained from all patients and also clearly states that the data only is available, and not any material. For the second cohort, Miao *et al.* state that informed consent was collect as well [33].

(3) Harm and benefit. Looking at genetic information can be helpful for research and diagnostic, but can also uncover mutations or other defects that are relevant for the patients but would not have been discovered otherwise. In the consent form, it must be made clear what the donor's or the patient's wish is for such cases.

(4) Good scientific conduct. Ideally, the scientific work should be conducted with the goal of helping the donor patients or any potential future patient. This work has this potential, since both breast cancer and immuno-oncology are broadly used therapies and both need more research. It can be stated though, that many patient do not have access to these expensive treatments and that this research will therefore only be helpful to a subset of the world's population.

Bibliography

[1] Seung Ho Shin, Ann M. Bode, and Zigang Dong. "Addressing the challenges of applying precision oncology". In: *NPJ precision oncology* 1.1 (2017), p. 28. ISSN: 2397-768X. DOI: 10.1038/s41698-017-0032-z.

[2] Hyuna Sung et al. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries". In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249. DOI: 10.3322/caac.21660.

[3] Andrew H. Sims et al. "Origins of breast cancer subtypes and therapeutic implications". In: *Nature clinical practice. Oncology* 4.9 (2007), pp. 516–525. DOI: 10.1038/ncponc0908.

[4] Carlos L. Arteaga et al. "Treatment of HER2-positive breast cancer: current status and future perspectives". In: *Nature reviews. Clinical oncology* 9.1 (2011), pp. 16–32. DOI: 10.1038/nrclinonc.2011.177.

[5] Allen M. Gown et al. "High concordance between immunohistochemistry and fluorescence in situ hybridization testing for HER2 status in breast cancer requires a normalized IHC scoring system". In: *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 21.10 (2008), pp. 1271–1277. ISSN: 0893-3952. DOI: 10.1038/modpathol.2008.83. URL: https://www.sciencedirect.com/science/article/pii/S0893395222023973.

[6] S. Ménard et al. "HER2 as a prognostic factor in breast cancer". In: *Oncology* 61 Suppl 2.Suppl. 2 (2001), pp. 67–72. ISSN: 0030-2414. DOI: 10.1159/000055404.

[7] M. Piccart et al. "The predictive value of HER2 in breast cancer". In: *Oncology* 61 Suppl 2.Suppl. 2 (2001), pp. 73–82. ISSN: 0030-2414. DOI: 10.1159/000055405.

[8] Jiani Wang and Binghe Xu. "Targeted therapeutic options and future perspectives for HER2-positive breast cancer". In: *Signal Transduction and Targeted Therapy* 4.1 (2019), p. 34. ISSN: 2059-3635. DOI: 10.1038/s41392-019-0069-2. URL: https://www.nature.com/articles/s41392-019-0069-2#citeas.

[9] Samaresh Sau and Arun K. Iyer. "PD-1 and PD-L1 Checkpoint Signaling Inhibition for Cancer Immunotherapy: Mechanism, Combinations, and Clinical Outcome". In: *Frontiers in Pharmacology* 8 (2017). DOI: 10.3389/fphar.2017.00561. URL: https://www.frontiersin.org/articles/10.3389/fphar.2017.00561.

[10] Alberto Carretero-González et al. "Analysis of response rate with ANTI PD1/PD-L1 monoclonal antibodies in advanced solid tumors: a meta-analysis of randomized clinical trials". In: *Oncotarget* 9.9 (2018), pp. 8706–8715. DOI: 10.18632/oncotarget.24283.

[11] Alexandra Snyder et al. "Genetic basis for clinical response to CTLA-4 blockade in melanoma". In: *The New England journal of medicine* 371.23 (2014), pp. 2189–2199. DOI: 10.1056/NEJMoa1406498.

[12] Eliezer M. van Allen et al. "Genomic correlates of response to CTLA-4 blockade in metastatic melanoma". In: *Science (New York, N.Y.)* 350.6257 (2015), pp. 207–211. DOI: 10.1126/science.aad0095.

[13] Edward B. Garon et al. "Pembrolizumab for the treatment of non-small-cell lung cancer". In: *The New England journal of medicine* 372.21 (2015), pp. 2018–2028. DOI: 10.1056/NEJMoa1501824.

[14] Pedro N. Aguiar et al. "The role of PD-L1 expression as a predictive biomarker in advanced non-small-cell lung cancer: a network meta-analysis". In: *Immunotherapy* 8.4 (2016), pp. 479–488. DOI: 10.2217/imt-2015-0002.

[15] Eric E. Walk et al. "The Cancer Immunotherapy Biomarker Testing Landscape". In: *Archives of pathology & laboratory medicine* 144.6 (2020), pp. 706–724. DOI: 10.5858/arpa.2018-0584-CP.

[16] Antonio C. Wolff et al. "HER2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update Summary". In: *Journal of oncology practice* 14.7 (2018), pp. 437–441. DOI: 10.1200/JOP.18.00206.

[17] Allen M. Gown. "Current issues in ER and HER2 testing by IHC in breast cancer". In: *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 21 Suppl 2 (2008), S8–S15. ISSN: 0893-3952. DOI: 10.1038/modpathol.2008.34.

[18] Soonmyung Paik et al. "Real-world performance of HER2 testing–National Surgical Adjuvant Breast and Bowel Project experience". In: *Journal of the National Cancer Institute* 94.11 (2002), pp. 852–854. ISSN: 0027-8874. DOI: 10.1093/jnci/94.11.852.

[19] Patrick C. Roche et al. "Concordance between local and central laboratory HER2 testing in the breast intergroup trial N9831". In: *Journal of the National Cancer Institute* 94.11 (2002), pp. 855–857. ISSN: 0027-8874. DOI: 10.1093/jnci/94.11.855.

[20] Daphne W. Bell. "Our changing view of the genomic landscape of cancer". In: *The Journal of pathology* 220.2 (2010), pp. 231–243. DOI: 10.1002/path.2645.

[21] Scott L. Carter et al. "Absolute quantification of somatic DNA alterations in human cancer". In: *Nature biotechnology* 30.5 (2012), pp. 413–421. DOI: 10.1038/nbt.2203.

[22] Peter van Loo et al. "Allele-specific copy number analysis of tumors". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.39 (2010), pp. 16910–16915. DOI: 10.1073/pnas.1009843107.

[23] Ana Conesa et al. "A survey of best practices for RNA-seq data analysis". In: *Genome biology* 17 (2016), p. 13. DOI: 10.1186/s13059-016-0881-8.

[24] Conor Giles Doran and Stephen R. Pennington. "Copy number alteration signatures as biomarkers in cancer: a review". In: *Biomarkers in medicine* 16.5 (2022), pp. 371–386. DOI: 10.2217/bmm-2021-0476.

[25] Geoff Macintyre et al. "Copy number signatures and mutational processes in ovarian carcinoma". In: *Nature Genetics* 50.9 (2018), pp. 1262–1270. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0179-8. URL: https://www.nature.com/articles/s41588-018-0179-8#citeas.

[26] Sander Ellegård et al. "ERBB2 and PTPN2 gene copy numbers as prognostic factors in HER2-positive metastatic breast cancer treated with trastuzumab". In: *Oncology letters* 17.3 (2019), pp. 3371–3381. ISSN: 1792-1074. DOI: 10.3892/ol.2019.9998.

[27] Zhihao Lu et al. "Tumor copy-number alterations predict response to immune-checkpoint-blockade in gastrointestinal cancer". In: *Journal for immunotherapy of cancer* 8.2 (2020). DOI: 10.1136/jitc-2019-000374.

[28] Dara S. Ross et al. "Next-Generation Assessment of Human Epidermal Growth Factor Receptor 2 (ERBB2) Amplification Status: Clinical Validation in the Context of a Hybrid Capture-Based, Comprehensive Solid Tumor Genomic Profiling Assay". In: *The Journal of molecular diagnostics : JMD* 19.2 (2017), pp. 244–254. DOI: 10.1016/j.jmoldx.2016.09.010.

[29] Stacey M. Stein et al. "Real-world association of HER2/ERBB2 concordance with trastuzumab clinical benefit in advanced esophagogastric cancer". In: *Future oncology (London, England)* 17.31 (2021), pp. 4101–4114. DOI: 10.2217/fon-2021-0203.

[30] Xin Shao et al. "Copy number variation is highly correlated with differential gene expression: a pan-cancer study". In: *BMC medical genetics* 20.1 (2019), p. 175. DOI: 10.1186/s12881-019-0909-5.

[31] Katherine A. Hoadley et al. "Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer". In: *Cell* 173.2 (2018), 291–304.e6. DOI: 10.1016/j.cell.2018.03.022.

[32] The Cancer Genome Atlas Network. "Comprehensive molecular portraits of human breast tumours". In: *Nature* 490.7418 (2012), pp. 61–70. DOI: 10.1038/nature11412.

[33] Diana Miao et al. "Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors". In: *Nature Genetics* 50.9 (2018), pp. 1271–1281. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0200-2.

[34] E. A. Eisenhauer et al. "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)". In: *European journal of cancer (Oxford, England : 1990)* 45.2 (2009), pp. 228–247. DOI: 10.1016/j.ejca.2008.10.026. URL: https://www.sciencedirect.com/science/article/pii/S0959804908008733.

[35] Raima Memon et al. "Discordance Between Immunohistochemistry and In Situ Hybridization to Detect HER2 Overexpression/Gene Amplification in Breast Cancer in the Modern Age: A Single Institution Experience and Pooled Literature Review Study". In: *Clinical Breast Cancer* 22.1 (2022), e123–e133. ISSN: 1526-8209. DOI: 10.1016/j.clbc.2021.05.004. URL: https://www.sciencedirect.com/science/article/pii/S1526820921001324.

[36] Shiuh-Wen Luoh et al. "HER-2 gene amplification in human breast cancer without concurrent HER-2 over-expression". In: *SpringerPlus* 2 (2013), p. 386. ISSN: 2193-1801. DOI: 10.1186/2193-1801-2-386.

[37] Amélie Boichard, Scott M. Lippman, and Razelle Kurzrock. "Therapeutic implications of cancer gene amplifications without mRNA overexpression: silence may not be golden". In: *Journal of hematology & oncology* 14.1 (2021), p. 201. DOI: 10.1186/s13045-021-01211-1.

[38] Ebru Sari et al. "Comparative study of the immunohistochemical detection of hormone receptor status and HER-2 expression in primary and paired recurrent/metastatic lesions of patients with breast cancer". In: *Medical Oncology* 28.1 (2011), pp. 57–63. ISSN: 1559-131X. DOI: 10.1007/s12032-010-9418-2. URL: https://link.springer.com/article/10.1007/s12032-010-9418-2#citeas.

[39] Huina Zhang et al. "Applying the New Guidelines of HER2 Testing in Breast Cancer". In: *Current oncology reports* 22.5 (2020), p. 51. DOI: 10.1007/s11912-020-0901-4.

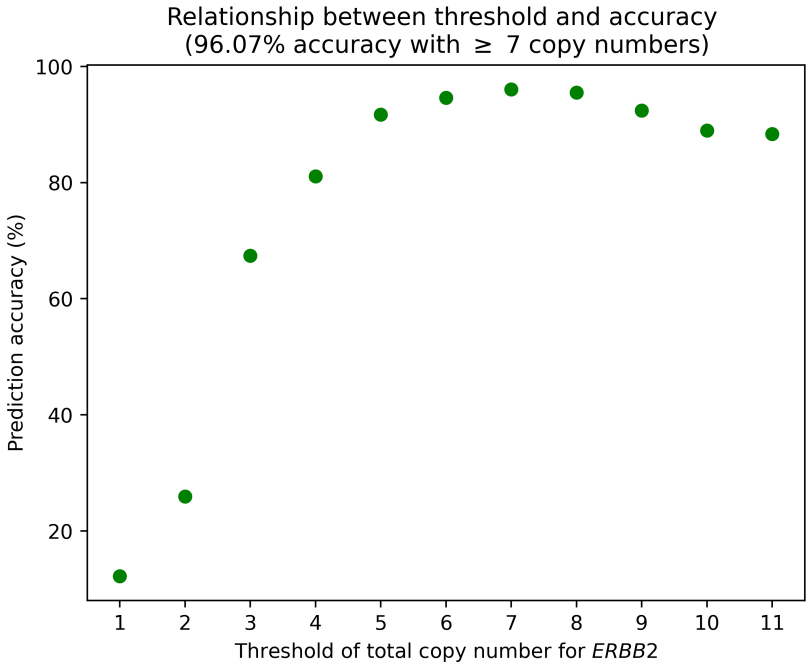
[40] Whijae Roh et al. "Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance". In: *Science translational medicine* 9.379 (2017). DOI: 10.1126/scitranslmed.aah3560.

[41] Tae Hee Hong et al. "Programmed Death-Ligand 1 Copy Number Alteration as an Adjunct Biomarker of Response to Immunotherapy in Advanced Non-small Cell Lung Cancer". In: *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* 0.0 (2023). DOI: 10.1016/j.jtho.2023.03.024. URL: https://www.jto.org/article/S1556-0864(23)00482-3/fulltext#%20.

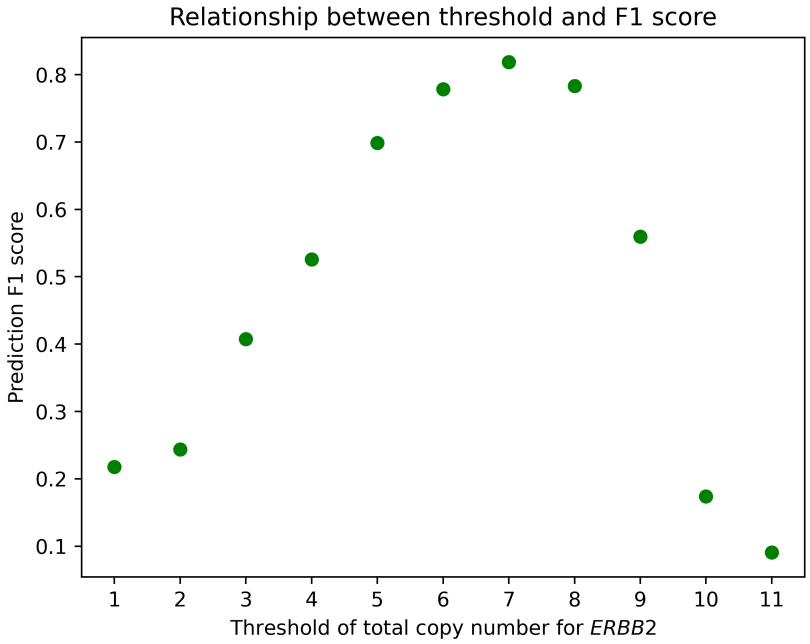
[42] Laura Morsberger et al. "HER2 amplification by next-generation sequencing to identify HER2-positive invasive breast cancer with negative HER2 immunohistochemistry". In: *Cancer Cell International* 22.1 (2022), p. 350. ISSN: 1475-2867. DOI: 10.1186/s12935-022-02761-1. URL: https://cancerci.biomedcentral.com/articles/10.1186/s12935-022-02761-1#citeas.

[43] Min Yan et al. "HER2 expression status in diverse cancers: review of results from 37,992 patients". In: *Cancer and Metastasis Reviews* 34.1 (2015), pp. 157–164. ISSN: 1573-7233. DOI: 10.1007/s10555-015-9552-6. URL: https://link.springer.com/article/10.1007/s10555-015-9552-6#citeas.

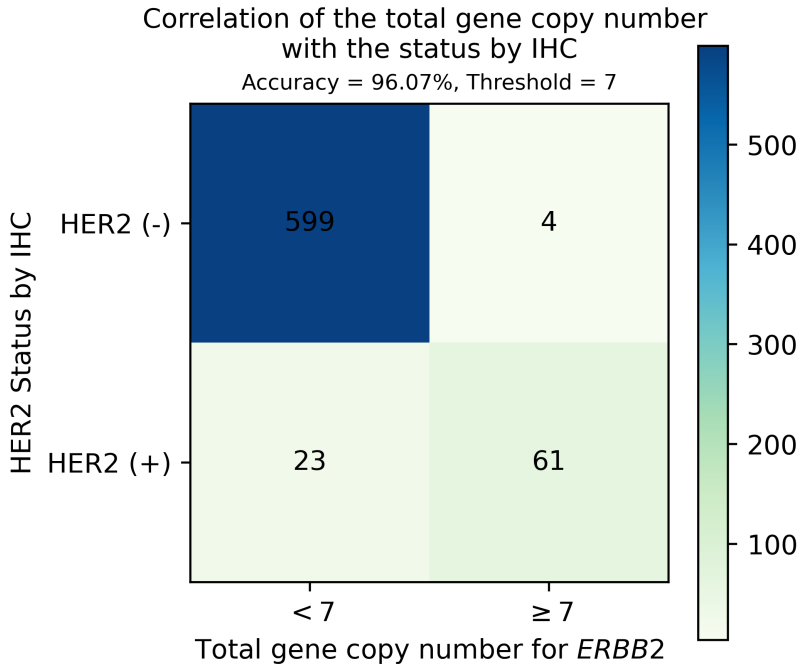
Supplementary Figures



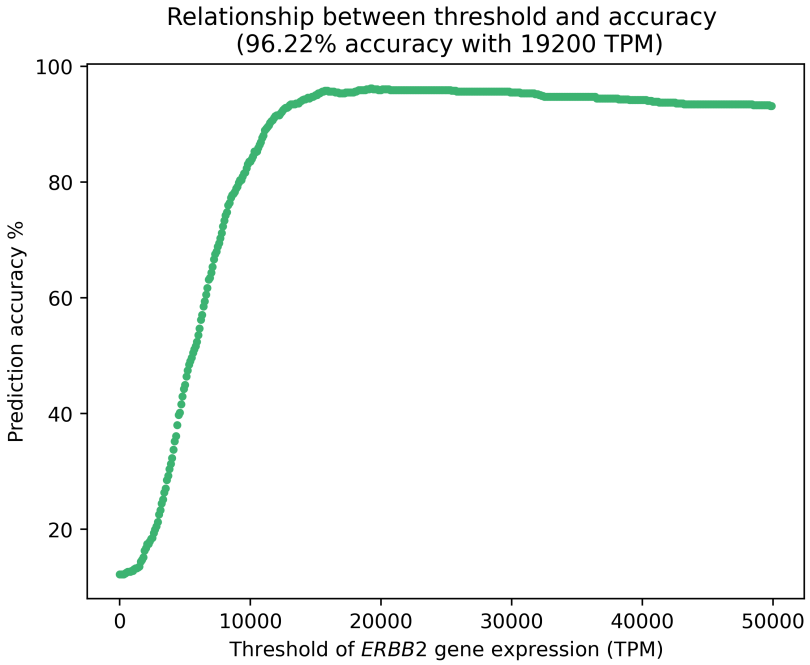
Supplementary Figure 1: Setting the threshold at different total *ERBB2* copy number resulted in these different accuracies. A threshold of 3 means that a sample with a copy number ≥ 3 is predicted as positive.



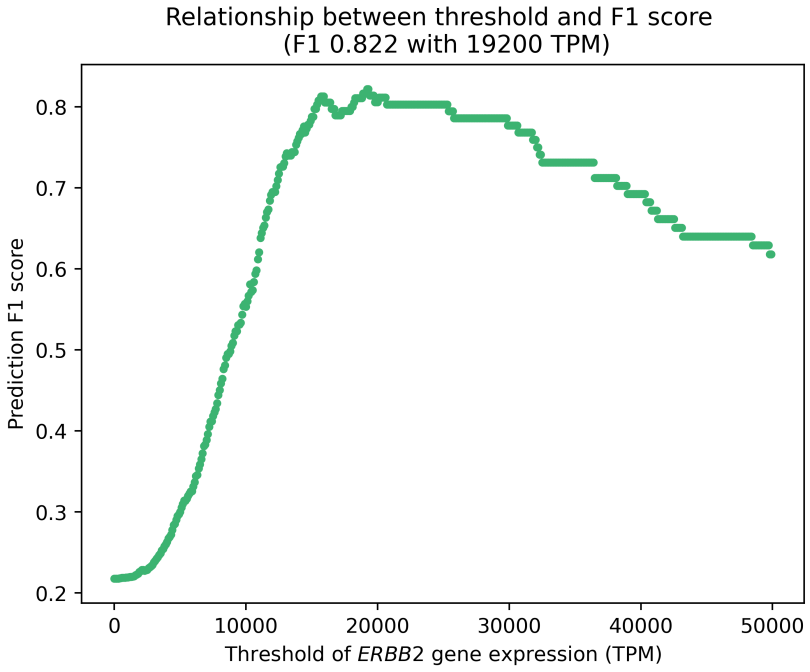
Supplementary Figure 2: Setting the threshold at different total *ERBB2* copy number resulted in these different F1 scores. A threshold of 3 means that a sample with a copy number ≥ 3 is predicted as positive. The F1 score accounted for the unbalance in positive and negative data.



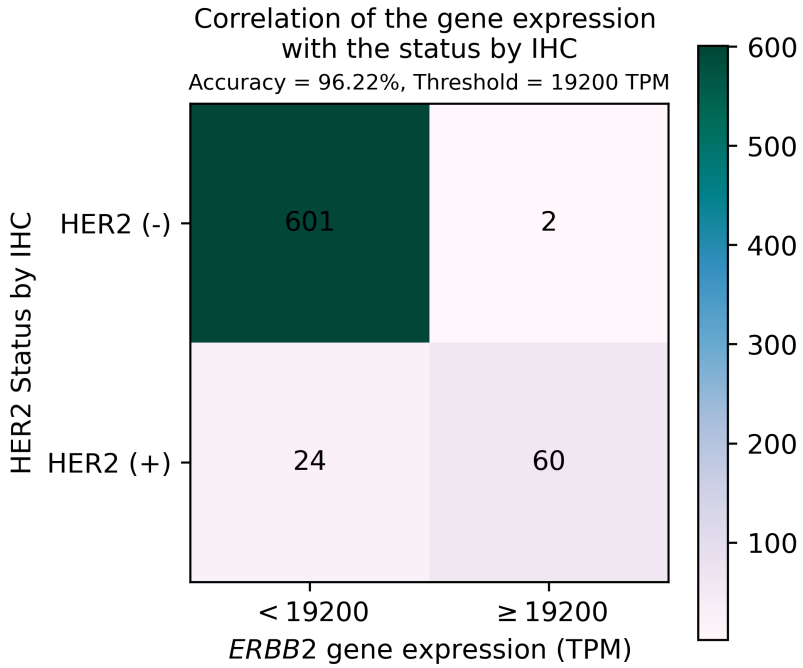
Supplementary Figure 3: A threshold set to 7 total *ERBB2* copy number resulted in this confusion matrix.



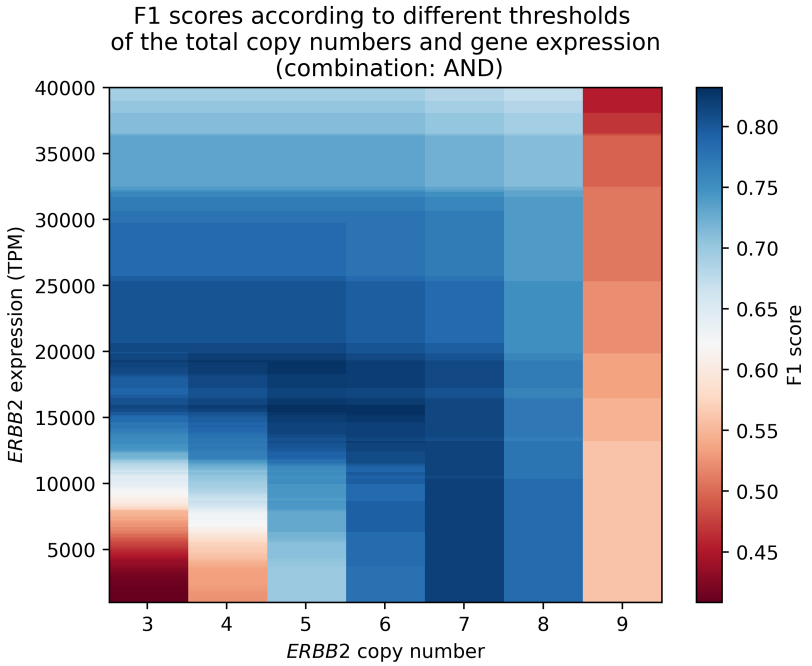
Supplementary Figure 4: Setting the threshold at different *ERBB2* expressions resulted in these different accuracies. A threshold of 10'000 means that a sample with a gene expression $\geq$ 10'000 is predicted as positive.



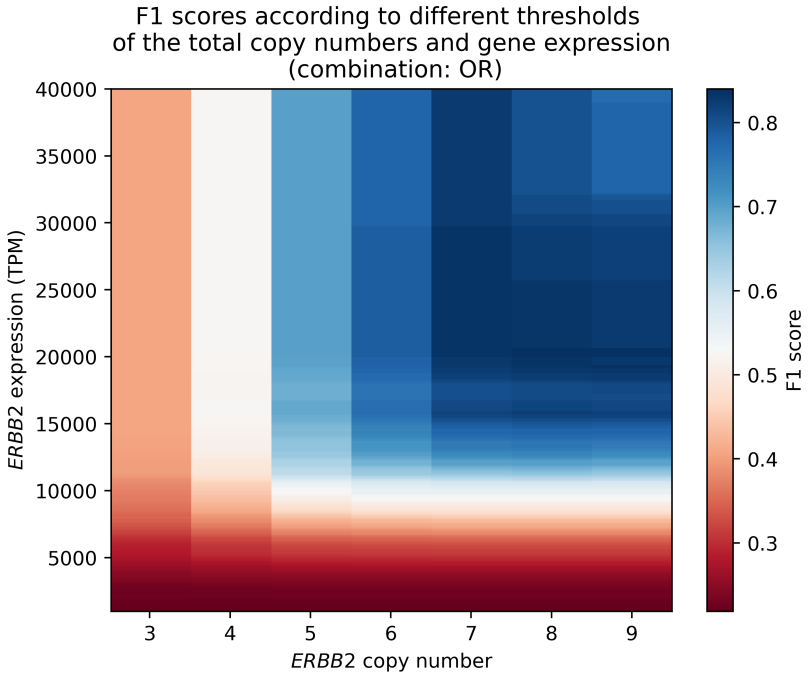
Supplementary Figure 5: Setting the threshold at different *ERBB2* expressions resulted in these different F1 scores. A threshold of 10'000 means that a sample with a gene expression $\geq$ 10'000 is predicted as positive. The F1 score accounted for the unbalance in positive and negative data.



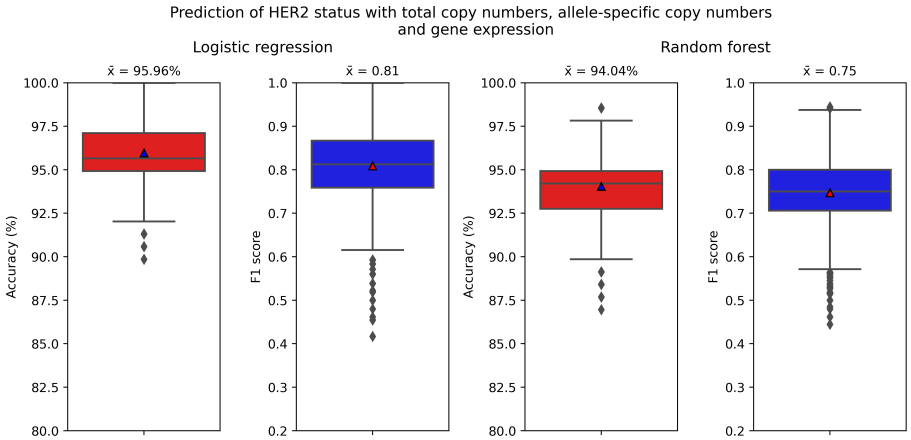
Supplementary Figure 6: A threshold set to 19'200 TPM for the *ERBB2* gene expression resulted in this confusion matrix.



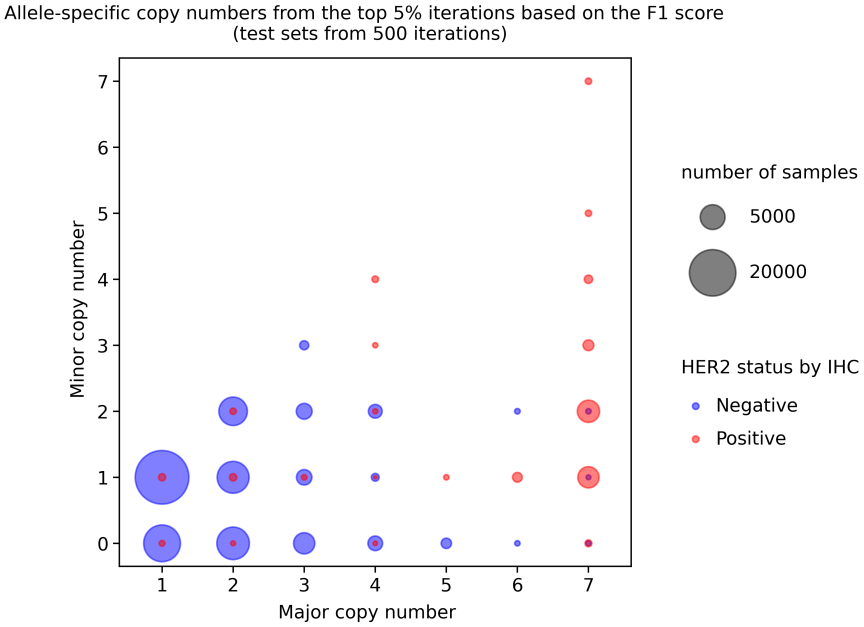
Supplementary Figure 7: F1 score in relationship with different combination of thresholds when integrating the two data modalities with a `AND` operator. This means both modalities need to be equal or higher than the set threshold for a given sample to be predicted as positive.



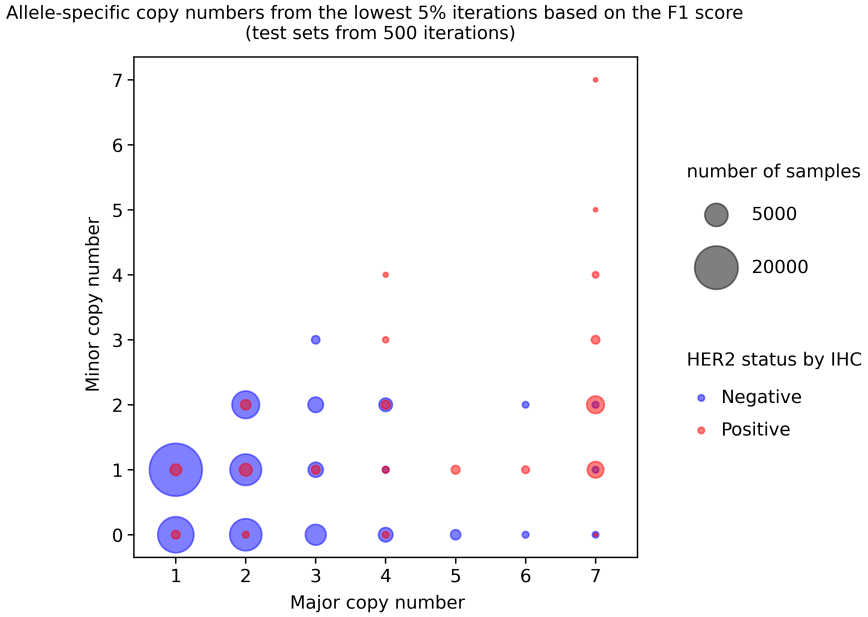
Supplementary Figure 8: F1 score in relationship with differen combination of threshold when integrating the two data modalities with a `OR` operator. This mean that samples with any modalities equal or higher than the set thresholds are predicted positive.



Supplementary Figure 9: Training a logistic regression model and a random forest model with the 687 samples and a 20:80 split. For each model type, 10'000 iterations of data splitting, training, predictions and assessment (accuracy and F1 score) have been made. Here, total copy number, allele-specific copy numbers and gene expression for *ERBB2* have been used. Triangles in boxplots are the means.



Supplementary Figure 10: After 10'000 iterations of logistic regression with the allele-specific copy numbers as features, a certain spread of the resulting F1 score has been observed (see figure 6, boxplot B). Shown here are the copy numbers of the 500 iterations that resulted in the highest F1 scores (0.909 and higher).



Supplementary Figure 11: After 10'000 iterations of logistic regression with the allele-specific copy numbers as features, a certain spread of the resulting F1 score has been observed (see figure 6, boxplot B). Shown here are the copy numbers of the 500 iterations that resulted in the lowest F1 scores (0.690 and lower).